

Simultaneous inference for functional data in sports biomechanics

Comparing statistical parametric mapping with interval-wise testing

Todd Colin Pataky · Konrad
Abramowicz · Dominik Liebl · Alessia
Pini · Sara Sjöstedt de Luna · Lina
Schelin

Received: date / Accepted: date

Abstract The recent sports science literature conveys a growing interest in robust statistical methods to analyze smooth, regularly-sampled functional data. This paper focuses on the inferential problem of identifying the parts of a functional domain where two population means differ. We considered four approaches recently used in sports science: interval-wise testing (IWT), statistical parametric mapping (SPM), statistical nonparametric mapping (SnPM) and the Benjamini-Hochberg (BH) procedure for false discovery control. We applied these procedures to both six representative sports science datasets, and also to systematically varied simulated datasets which replicated ten signal- and/or noise-relevant parameters that were identified in the experimental datasets. We observed generally higher IWT and BH sensitivity for five of the six experimental datasets. BH was the most sensitive procedure in simulation, but also involved relatively high false positive rates (generally >0.1) which increased sharply (>0.3) in certain extreme simulation scenarios

T. C. Pataky
Department of Human Health Sciences, Kyoto University, Kyoto, Japan
Tel.: +81 (0)75 751 4175
E-mail: pataky.todd.2m@kyoto-u.ac.jp

K. Abramowicz
Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden

D. Liebl
Department of Statistics, University of Bonn, Bonn, Germany

A. Pini
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy

S. Sjöstedt de Luna
Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden

L. Schelin
Unit of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden

including especially highly rough data. SPM and SnPM were more sensitive than IWT in simulation except for (i) very rough data, (ii) high nonstationarity, and (iii) highly nonuniform smoothness. These results suggest that the optimum procedure is both signal and noise-dependent. We conclude that: (i) BH is most sensitive but also susceptible to high false positive rates, (ii) IWT, SPM and SnPM appear to have relatively inconsequential differences in terms of domain identification sensitivity, except in cases of extreme signal/noise characteristics, where IWT appears to be superior at identifying a greater portion of the true signal.

Keywords one-dimensional functional data · local inference · continuum data analysis · simulation · signal modeling · kinematics · biomechanics

1 Introduction

Biomechanics is a sports science-related field of research that applies mechanical engineering principles to the study of biological motion. Biomechanical data are often manifested as observations of one-dimensional (1D) functional data which represent the often highly smooth change of a (bio)mechanical parameter over a temporal or spatial domain. Representative and motivating examples include external reaction force changes over the pedaling cycle in cycling (Kautz et al., 1991) (Fig.1), joint angle changes over time during stance phase in sports manoeuvres (Neptune et al., 1999) (Fig.2), and muscle force changes over the gait cycle (Besier et al., 2009) (Fig.3).

One-dimensional biomechanical data are almost always continuous and are almost always sampled regularly over the domain. The main reason is that the underlying measurement equipment, including for example force plates and motion capture systems, have firmly engineered, hardware-based sampling rates of 200 Hz or greater, and the total temporal duration of most routinely assessed activities exceeds 500 ms. The tendency is to over-sample the original data to ensure that subsequent downsampling does not destroy the original signal. The resulting data are often highly smooth, even when resampled to registered domains of approximately 100 points (Figs.1-3).

The sports science literature contains inconsistent terminology regarding these data. They have been referred to as, for example: ‘functional data’, ‘curves’, ‘trajectories’, ‘waveforms’, ‘time series’, ‘continua’, and ‘1-D data’. Regardless of terminology, historically the most common data analysis approach in the sports science literature has been to further reduce the data by extracting scalars from a small subset of often pre-defined events. For example, joint angles at stance initiation, or muscle forces at the instant of maximal joint flexion. This scalar extraction procedure unsurprisingly results in a loss of information (Richter et al., 2014; Pataky et al., 2015) due primarily to under-sampling and thereby a violation of the Nyquist sampling criterion (Shannon, 1949). Recently, it has become increasingly common to analyze the whole dataset instead, without further reduction (Dannenmaier et al., 2020; Warmenhoven et al., 2018).

The main aim of this paper is to highlight the similarities and differences between four simultaneous inference procedures that have recently been used to analyze the aforementioned class of data: data whose dependent variable's values vary smoothly over a regularly sampled 1D domain. The four procedures — Interval-Wise Testing (IWT) (Pini and Vantini, 2017), Statistical Parametric Mapping (SPM) (Pataky et al., 2013), Statistical non-Parametric Mapping (SPM) (Pataky et al., 2015) or equivalently 'Fmax' (Ramsay and Silverman, 2005), and the Benjamini-Hochberg (BH) procedure for false discovery rate (FDR) control (Benjamini and Hochberg, 1995; Naouma and Pataky, 2019; Olsen et al., 2021) — are compared in this paper both using publicly available datasets from the sports science literature, and in simulation under a relatively wide range of signal and noise scenarios. The focus of the comparisons are hypothesis testing and identification of parts of the domain where significant differences occur.

We selected these four procedures because (i) they have all been used in sports biomechanics research, yet are rarely or never all used in single studies, and (ii) it is unclear how sports biomechanics results may be affected by the choice of inferential procedures. Through a relatively thorough comparison of these common methods, we intend to identify advantages and disadvantages and to provide guidance regarding the situations and dataset characteristics for which the choice of method may affect the results.

For simplicity and in interest of manuscript conciseness, this paper limits focus to the two-population inferential problem, whose aim is to identify the parts of the domain where two functional population means differ significantly from each other. One of the main issues of functional inferential procedures is multiplicity. Since whole functions are analyzed, the data domain is theoretically an infinite dimensional space. The inferential problem under consideration hence leads to a multiple hypothesis testing problem that needs to be addressed, and all four procedures offer solutions to this problem.

SPM solves the multiple testing problem in a relatively simple manner, using a small number of parameters, and the known topological behavior of Gaussian random fields (Adler and Taylor, 2007), and specifically the expected Euler characteristic of a thresholded domain, to parametrically calculate critical test statistic thresholds. The parameters include just (i) the usual degrees of freedom parameter, and (ii) a smoothness parameter which, for the case of a one-dimensional domain, reduces to a single scalar (Friston et al., 2007). The smoothness parameter can be estimated from functional residuals (Kiebel et al., 1999), and is linearly related to the residuals' Lipschitz-Killing curvature (LKC). This is a relatively simple, computationally efficient approach that has been shown to be robust to a variety of real sports biomechanics data distributions (Pataky et al., 2015) and also to a variety of simulated data imperfections, including non-uniform smoothness across the domain (Pataky et al., 2019).

SnPM (Nichols and Holmes, 2002) (or 'Fmax') solves the problem in the identical manner, but typically uses observation label permutation to build the null distribution non-parametrically. It has been shown that SnPM solu-

tion numerically converges to the Gaussian random field solution when the data are normally distributed, and when the sample size is sufficiently large to afford a reasonably dense permutation distribution (Pataky, 2016). While it has also been shown that the choice of SPM vs. SnPM has little-to-no practical consequences for example biomechanical datasets (Pataky et al., 2015; Warmenhoven et al., 2018), these procedures have previously not been systematically compared across a variety of datasets or in simulation.

IWT is also an inferential procedure for functional data, which also corrects for the multiple testing problem, with the specific aim of selecting intervals of the domain where the null hypothesis is violated (Pini and Vantini, 2017). The multiplicity issue is handled by the definition of an adjusted p-value function that can be used for domain selection. In its original formulation, IWT is based on nonparametric permutation tests, making it computationally less efficient than SPM, but also freeing it from distributional and smoothness assumptions. Like SPM, IWT has previously been applied to a variety of sports related movement data (Hébert-Losier et al., 2018; Markström et al., 2019).

The BH procedure for FDR control Benjamini and Hochberg (1995) is a versatile multiplicity correction procedure that has been widely employed for the analysis of 3D/4D functional data in the Neuroimaging field (Genovese et al., 2002; Chumbley and Friston, 2009). While the BH procedure has also recently been applied to 1D biomechanical data analysis (Naouma and Pataky, 2019) — and also more extensively in simulation to 1D functional data (Olsen et al., 2021) — they are not routinely applied to functional data analysis in the sports science literature.

The paper is organized in the following manner. First, in Section 2 we summarize the four methods (IWT, SPM, SnPM and BH) in the context of subsequently described simulations. In Section 3 we compare these methods using experimental data, as motivation for simulation scenarios we subsequently consider. In Section 4 we describe a simulation study in detail, including a total of ten parameters relevant to signal and error modeling. A discussion is provided in Section 5 which considers the implications of the results in the context of both the statistics and sports science literatures.

2 Method

Let $f_{1i}(t), t \in \mathcal{D}, i = 1, \dots, n_1$ be a sample of independent functional observations from a population with mean $\mu_A(t)$ and let $f_{2i}(t), t \in \mathcal{D}, i = 1, \dots, n_2$ be a sample of independent functional observations from another population with mean $\mu_B(t)$. Assume that all functions are observed on a common domain $\mathcal{D} = [a, b] \subset \mathbb{R}$, where we set, without loss of generality, $\mathcal{D} = [a, b] = [0, 1]$ for the remainder of the paper. We want to perform tests of the following hypotheses:

$$\begin{aligned} H_0 : \mu_A(t) &= \mu_B(t), \forall t \in \mathcal{D} \\ H_1 : \exists t \in \mathcal{D} \text{ s.t. } \mu_A(t) &\neq \mu_B(t). \end{aligned} \tag{1}$$

Test (1) is however only global: a rejection of H_0 would only imply that there exists at least one point in $\mathcal{D}$ where the two populations differ in mean, but it would not be possible from an H_0 rejection to select the portions of the domain where H_0 is violated. To select domain portions, we instead focus on the following family of local tests, indexed by $t \in \mathcal{D}$:

$$\begin{aligned} H_0^t &: \mu_A(t) = \mu_B(t) \\ H_1^t &: \mu_A(t) \neq \mu_B(t). \end{aligned} \quad (2)$$

Hypotheses (2) can be tested point-wise, and a p-value $p(t)$ can be computed (with parametric or nonparametric methods) at each $t \in \mathcal{D}$. Together, point-wise p-values make up a function $p : T \rightarrow [0; 1]$, which can be termed the ‘unadjusted’ p-value function (Pini and Vantini, 2017). Clearly, the unadjusted p-value function is not suitable for selecting the parts of the domain where the null hypothesis is violated, since there is no adjustment for multiplicity across domain points. This paper therefore focuses on techniques for adjusting the p-value function to control the family-wise error rate (FWER), that is the probability of committing at least one type I error over the domain. Both SPM and IWT share this FWER control property.

2.1 Statistical parametric mapping (SPM)

SPM was developed in the applied Neuroimaging literature, primarily for the analysis of 3D fMRI brain images and image time series (Friston et al., 1995, 2007). In the context of classical hypothesis testing, SPM uses Gaussian random field theory (Adler and Taylor, 2007) to extend classical procedures to the domain of n -manifolds, for which parametric solutions exist for expected topological characteristics, including especially the expected Euler characteristic (EEC) (Adler and Hasofer, 1976). The EEC for Gaussian functional data is given by (Pataky, 2016) as:

$$P(\exists t \in [0, 1] : T(t) \geq u) \leq EEC(u) \quad (3)$$

where:

$$EEC(u) = \frac{S}{W} \frac{\sqrt{4 \log 2}}{2\pi} e^{-u^2/2} \quad (4)$$

and where $T(t)$ denotes the functional t test statistic, u is an arbitrary threshold, S is the domain size (1.0 in this paper), and W is the full-width-at-half-maximum (FWHM) of a Gaussian kernel which, when convolved with uncorrelated, Gaussian functional data, would yield the same Lipschitz-Killing curvature (LKC) as the observed functional residuals; the FWHM is linearly related to both the LKC and the smoothness parameter of the Matérn covariance function. This EEC can be easily extended to arbitrary functional test statistics including the t statistic, which in turn yield parametric survival functions (Worsley et al., 1996) from which adjusted p-value functions can

be calculated. We note that these survival functions are valid generally only for high thresholds (approximately p-value < 0.5) (Friston et al., 2007); since this paper is only interested in low p-values (near α), this inaccuracy for high p-values was not considered problematic.

More conceptually, since (4) shows that the EEC is directly affected by the domain smoothness (i.e., the FWHM), and since the EEC is directly related to the null hypothesis (1), it is clear that the EEC can be used to implement a smoothness-dependent correction for multiple comparisons across all domain points. This multiple comparisons correction converges to standard test statistic distributions for large FWHM, approaching infinity, involving functions of the form $f(x) = c$, where c is a constant.

While SPM is primarily used in the applied literature for 3D brain image analysis, a 1D software implementation with a simplified user interface (Pataky, 2012a) has seen SPM branch into other applied fields like Sports Science, especially for the analysis of biomechanical data including kinematics and force trajectories (Pataky et al., 2013). SPM’s underlying topological expectations and their utility for parametric, classical hypothesis testing involving functional data have been numerically validated elsewhere (Pataky, 2016). SPM routines assume Gaussian-distributed data at each domain point, and also assume uniform smoothness across the domain. The latter assumption has nevertheless been shown to be irrelevant to H_0 rejection decisions (Pataky et al., 2019). SPM probabilities are based on continuous domains and continuous domain derivatives, but software implementations deal predominantly with discretized data and discrete derivative approximations. SPM is sensitive to discretization resolution only when the resolution becomes coarse with respect to the signal’s frequency characteristics (Pataky, 2012b).

2.2 Statistical non-parametric mapping (SnPM)

As pointed out in Ramsay and Silverman (2005, Ch.13.3.3), one nonparametric approach to hypothesis testing for functional data is label permutation. Methods involving the permutation distribution of the maximum test statistic have been shown to converge to SPM’s parametric EEC solutions when the underlying data are Gaussian (Nichols and Holmes, 2002; Pataky et al., 2015). Permutation and related nonparametric techniques are generally referred to as ‘SnPM’ in the applied literature (Nichols and Holmes, 2002); we retain the term ‘SnPM’ in this paper due to its SPM convergence properties. Using this approach, adjusted p-value functions can be calculated simply as percentiles from the permutation distribution.

It has also been shown that numerical differences between SPM and SnPM critical thresholds are typically inconsequential for the analysis of experimental data in sports science (Pataky et al., 2015). To our knowledge, SPM’s and SnPM’s critical thresholds have not been systematically explored in simulation.

2.3 Interval-wise testing (IWT)

The IWT was introduced by Pini and Vantini (2017) as a method that deals with domain selection, while correcting for the problem of multiplicity. The IWT was introduced in the context of comparing the mean functions between two populations, but have e.g., been extended to linear models (Abramowicz et al., 2018) and to a multi-aspect setting (Pini et al., 2019).

The method is based on testing the hypothesis of equality in mean functions on every interval $\mathcal{I}$ of the domain $\mathcal{D}$ using a desired test statistic. From these tests, using the supremum over all tests on intervals, an adjusted p-value function ($\tilde{p}(t), \forall t \in \mathcal{D}$) is derived to control the interval-wise error rate (Pini and Vantini, 2017). It can be thresholded and used for domain selection purposes, and we then have a control of the probability of falsely detecting intervals with significant group differences. The obtained control is weaker than the control of the family wise error rate (FWER), and the probability of making at least one type-I error is controlled only if H_0 is true on an interval. If the null hypothesis is true on e.g. a union of intervals, IWT fails to control the FWER.

The IWT procedure, applied to integrated point-wise test statistics, was implemented using non-parametric permutation tests, resulting in a strategy that does not require the specification of any parametric distributional model of functional data. However, the method is not limited to such non-specified tests. In practice, for implementation, the domain must be discretized to a sufficiently fine resolution, see Pini and Vantini (2017).

2.4 Benjamini-Hochberg (BH) procedure

FDR control was introduced by Benjamini and Hochberg (1995) as a versatile multiple testing correction procedure. When simultaneously conducting multiple tests, the FDR is defined as the proportion of incorrectly rejected null hypotheses amongst all rejected null hypotheses. For example, if 20 tests of 100 were rejected and two of those were incorrectly rejected, the FDR is 0.1. The originally proposed FDR control procedures of Benjamini and Hochberg (1995) are readily adaptable to both weighted and functional measures (Olsen et al., 2021).

3 Experimental data

SPM, SnPM and BH analyses of experimental data were conducted in Python 3.8.3, (van Rossum, 2014) using Anaconda 3.6.10 (Anaconda, 2020), and IWT analyses were conducted in R 4.0.3 (R Core Team, 2020). R and Python code, along with experimental data, are provided as supplementary material in a public repository (<https://github.com/0todd0000/iwtspm>).

3.1 Dataset descriptions

Six publicly available datasets related to sport science (Table 1; Figs.1-4, left panels) were used to compare IWT, SPM, SnPM and BH. Two-tailed tests were performed for each dataset using each of the three methods. For all datasets we aimed to identify the domain regions where the two functional population means exhibited significant differences. A significance level of $\alpha = 0.05$ was used throughout. We note that all experimental datasets exhibited both substantial non-stationarity and substantial non-uniformity (see Appendix, Figs.12 - 13), and that the Besier et al. (2009) datasets exhibited substantial non-normality (Appendix, Fig.14).

Table 1 Dataset descriptions. Six datasets (i.e., six different dependent variables) from four data repositories were analyzed.

Data repository & Task	Group	n_1, n_2	DVs	Unit	Data resolution
Kautz et al. (1991); Cycling	(1) Moderate (2) High workload	10,10	(a) Normal and (b) Tangential pedal forces	N	101
Neptune et al. (1999); Cutting movements	(1) Side shuffle (2) V-cut	8,8	Knee flexion	Deg	101
Besier et al. (2009); Walking	(1) Healthy (2) PFP	15,26	(a) Medial gastrocnemius and (b) Semimembranosus forces	N	100
Dorn et al. (2012); Running	(1) Running (2) Sprinting	4,4	Anterior ground reaction force (GRF)	N	100

Deg: Degrees; DV: Dependent variable; N: Newton; PFP = Patellofemoral pain

3.2 Results

Hypothesis testing results are presented as adjusted p-value functions in the right panels of Figs. 1-4. Unadjusted p-value functions are provided in these figures as a reference, representing point-wise test results which are not adjusted for multiple testing.

BH and IWT generally exhibited lower p-values, with sub- α p-values across larger portions of the domain than SPM and SnPM (Fig. 1-4). The only notable exception was the Besier et al. (2009)(a) dataset, for which both SPM and SnPM exhibited smaller p-values over the approximate domain region: 0.25 - 0.35 (Fig.3, top right panel). Null hypothesis rejection decisions (i.e., whether the minimum, adjusted p-value was smaller than the predefined significance level of $\alpha=0.05$), were generally identical for all procedures, except for the Neptune et al. (1999) dataset, for which IWT yielded significance ($p_{\min} = 0.049$) while SPM, SnPM and BH did not ($p_{\min} = 0.085, 0.086$ and 0.069 , respectively) (Fig. 2).

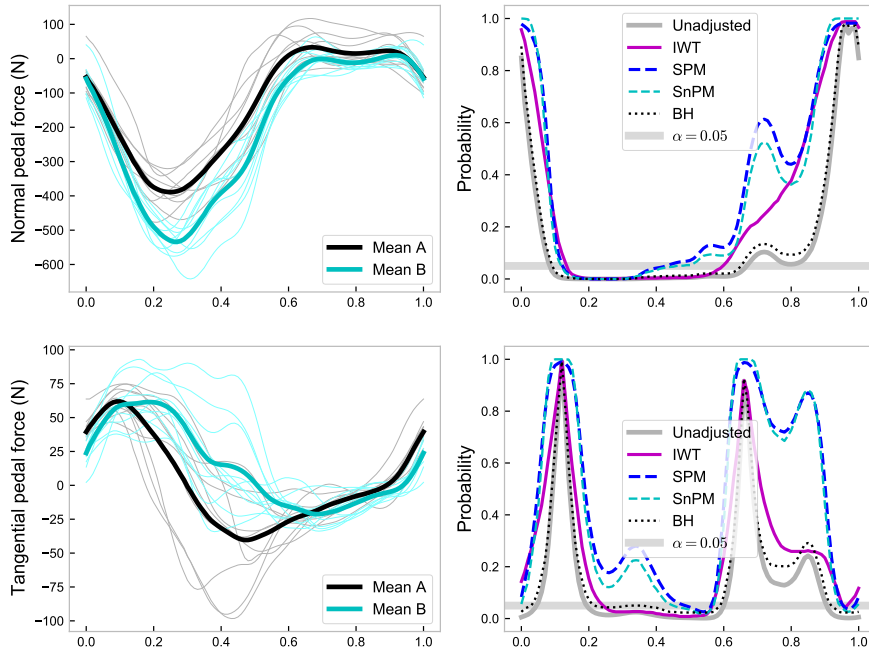


Fig. 1 Results for normal (top panels) and tangential pedal force (bottom panels) from [Kautz et al. \(1991\)](#), see also Table 1. Left panels: data and sample means. Right panels: corresponding p-value functions with threshold level 0.05.

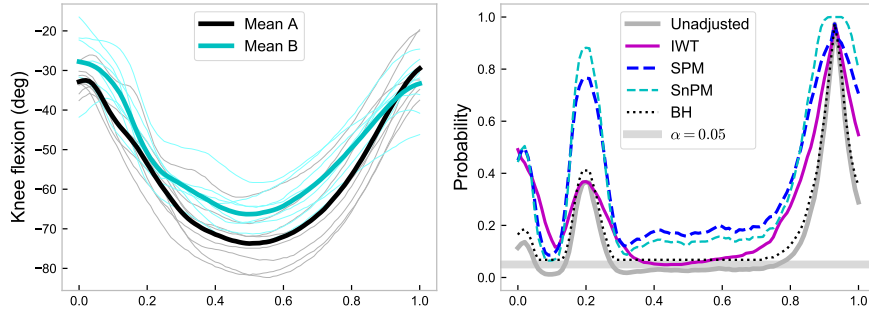


Fig. 2 Results for the [Neptune et al. \(1999\)](#) knee flexion dataset; data presented as in Fig. 1.

3.3 Summary

The experimental datasets exhibited a wide variety of smoothness, variance, and apparent effect characteristics including: broad effects (e.g. Fig. 1), multiple areas of relatively large effects (e.g. Fig. 4), relatively large variance in large-effect regions (Fig. 1-bottom), and relatively large smoothness in large-effect regions (e.g. Fig. 2). Additionally, the different inferential procedures exhibited

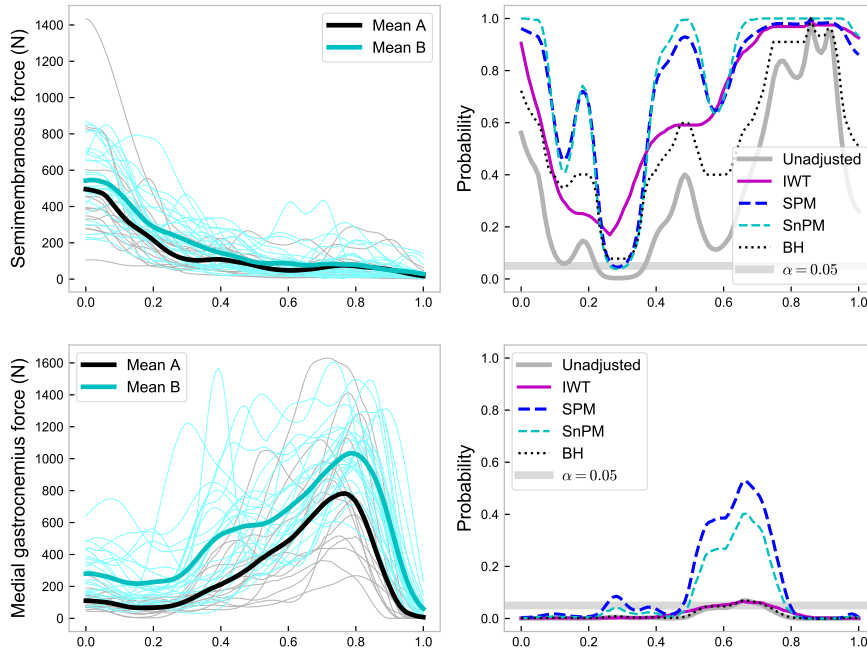


Fig. 3 Results for semimembranosus (top panels) and medial gastrocnemius muscle forces (bottom panels) from [Besier et al. \(2009\)](#); data presented as in Fig.1.

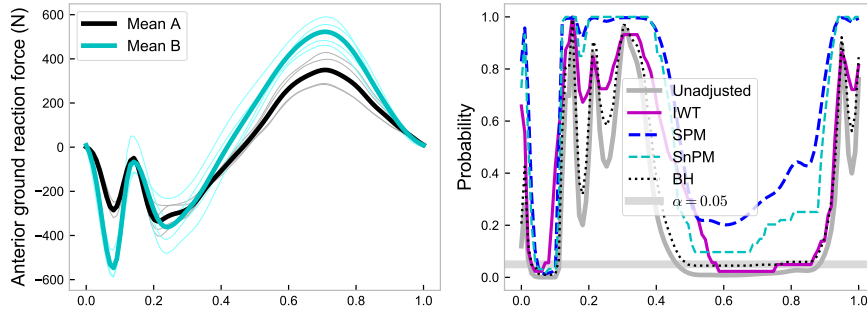


Fig. 4 Results for the [Dorn et al. \(2012\)](#) running force dataset; data presented as in Fig.1.

a variety of trends including (a) relatively small IWT p-values in broad large-effect regions (e.g. Fig.1-bottom) and (b) the opposite in narrower large-effect regions (Fig.3-top). Since it is unknown whether the aforementioned dataset characteristics are systematically related to the relative performances of the different inferential procedures, we conducted a simulation study to emulate these key dataset features, and also to systematically manipulate their magnitudes, with the ultimate goal of elucidating how particular dataset features are related to the different procedures' performances.

4 Simulation Study

4.1 Simulation settings

We denote $\mathcal{D}_0 = \{t \in \mathcal{D} : H_0^t \text{ is true}\}$ the part of the domain $\mathcal{D}$ where the null hypothesis is true, and $\mathcal{D}_1 = \{t \in \mathcal{D} : H_0^t \text{ is false}\}$ the part of the domain where the null hypothesis is false (i.e., the part of the domain with true signal), see Fig. 5.

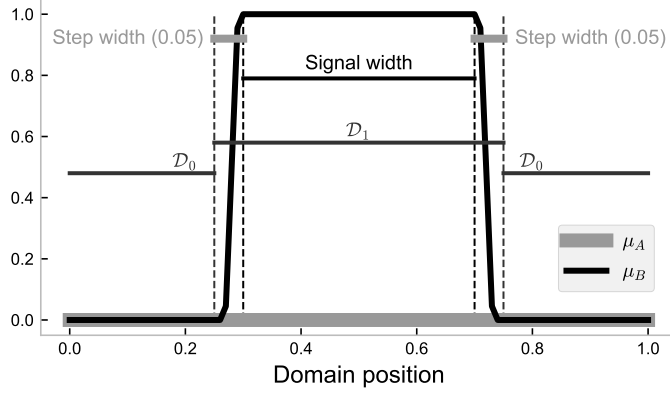


Fig. 5 Baseline signal model. Sets $\mathcal{D}_0$ and $\mathcal{D}_1$ depict the domain regions that correspond to truth of H_0 and H_1 , respectively. In simulations, ‘signal width’ was defined as the plateau width (0.4 in the depicted baseline case), but $\mathcal{D}_1$ spanned both the plateau and sigmoid steps on both sides as depicted; the sigmoid step width was 0.05 in all simulations, irrespective of signal width.

A baseline simulation scenario was defined with a single signal region (i.e., where $\mathcal{D}_1$ was a single interval; Fig 5), and where simulation parameters (Table 2), including a default standard deviation of 1.0 for both populations, were chosen to qualitatively reproduce the main features of the experimental data results above. An example realization of the baseline scenario is depicted in Fig. 6. Simulations explored the effects of changing one parameter at a time relative to baseline. All tested parameter values are listed in Table 2. For each set of parameters, 10,000 simulated datasets were generated. For all simulations, the domain was discretized using 101 points.

Two sample size-relevant parameters were investigated: sample size, and sample imbalance. Sample size was defined as the number of functional observations in each of the two groups (‘A’ and ‘B’), and the baseline scenario was $n_A = n_B = 20$. Sample imbalance was introduced by setting n_A between 4 and 36 then and maintaining $n_A + n_B = 40$.

Three signal-relevant parameters were investigated including: (i) signal amplitude, (ii) signal width (see Fig. 5), and (iii) multi-signal pulse width (not part of the baseline scenario). The latter parameter involved a scenario with

Table 2 Simulation parameters. Baseline scenario parameters are listed in bold.

Category	Parameters	Baseline		
Sample	¹ Sample size (n_A)	5, 10, 15	20	25, 30, 40, 50, 80
	¹ Imbalance (n_A)	4, 6, 8, ...	20	..., 32, 34, 36
Signal	² Amplitude	0, 0.5	1	1.5, 2, 2.5, 3
	Width	0.05, 0.1, 0.2, 0.3	0.4	0.5, 0.6
	³ Multi-pulse width	0.04, 0.08, 0.12		0.16, 0.20
Smoothness	⁴ FWHM	0.03, 0.1	0.2	0.3, 0.4, 0.5, 0.6
	⁵ FWHM ratio $\mathcal{D}_0/\mathcal{D}_1$	$\frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}$	1	2, 3, 4, 5
Variance	⁶ σ ratio $\mathcal{D}_0/\mathcal{D}_1$	$\frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}$	1	2, 3, 4, 5
	⁶ Unequal var (σ_A/σ_B)	$\frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}$	1	2, 3, 4, 5
	⁷ Skewness (α)	-5, -4, -3, -2, -1	0	1, 2, 3, 4, 5

1. n_A and n_B are group sample sizes; for the ‘sample size’ simulations: $n_A = n_B$, and for the ‘imbalance’ simulations: $n_A + n_B = 40$.
2. Amplitude=0 represents a null signal, for which $\mathcal{D}_1$ is the null interval.
3. Multi-pulse signals (Fig.7) were not included in the baseline scenario.
4. FWHM: full width at half maximum (smoothness parameter).
5. $\mathcal{D}_0$: the domain where H_0 is true; $\mathcal{D}_1$: the domain where H_1 is true;
6. σ : standard deviation. Group B variance was constant ($\sigma_B^2 = 1$) for unequal variance simulations.
7. Skewness parameter α refers to the standard skew normal distribution.

three signal regions, i.e., where $\mathcal{D}_1$ was a union of three intervals centered on positions 0.2, 0.5, and 0.8 (Fig.7). The multi-signal pulse width was varied systematically with values: $\{0.04, 0.08, 0.12, 0.16, 0.20\}$, where each width applied to all three pulses simultaneously. Since these multiple pulses were also modeled with a sigmoid step of 0.05 width (Fig.5), in this multi-pulse scenario $\mathcal{D}_1$ spanned a total domain proportion of $3(0.05 + w + 0.05)$, where w is the pulse width, so the minimum and maximum sizes of $\mathcal{D}_1$ were 0.42 and 0.90, respectively.

Two smoothness parameters were investigated: (i) Matérn smoothness, here described using the FWHM parameter (Kiebel et al., 1999), with higher FWHM values corresponding to smoother functions, and (ii) FWHM ratio. The FWHM ratio parameter controlled the smoothness and variance over $\mathcal{D}_0$ and $\mathcal{D}_1$ separately (Fig.8), and was defined relative to the signal region $\mathcal{D}_1$, where ratios greater and less than one imply larger and smaller values in $\mathcal{D}_1$, respectively. Nonuniformly smooth error functions were generated using previously described code (Yang and Ren, 2019, Section 2.1).

Three variance parameters were investigated: SD ratio, unequal variance, and skewness. The SD ratio parameter, like the smoothness parameter above, was defined relative to $\mathcal{D}_1$, whose SD was held at 1 to maintain a constant true effect size. Unequal variance was defined in terms of the two groups (σ_A^2 and

σ_B^2); σ_B was held at a constant value of 1, like the SD ratio case, to maintain a constant true effect size, and σ_A was varied between 0.2 and 5.

Skewness was simulated using the skew normal distribution, where a skewness of 0 corresponds to a Gaussian error distribution (see Appendix, Fig. 15 for an example skewed dataset).

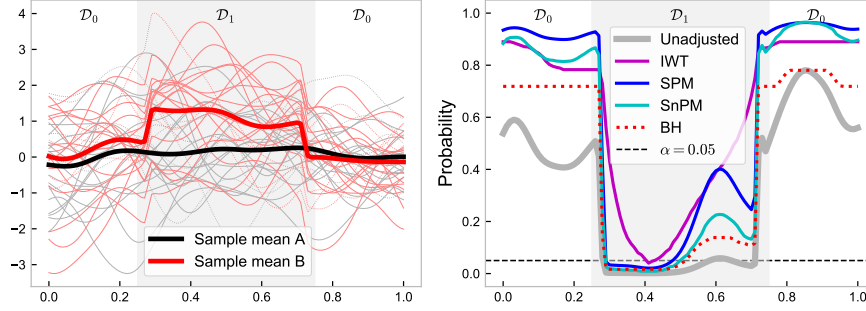


Fig. 6 Example realization of the baseline scenario (see Table 2 for simulation parameter values). Left panel: data including sample means. Right panel: corresponding p-value functions with threshold level 0.05.

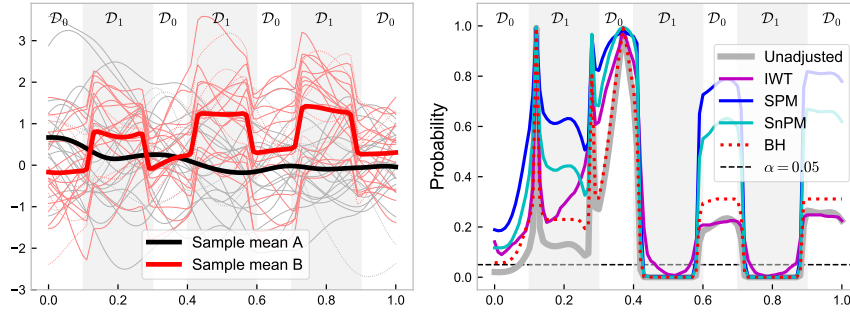


Fig. 7 Example realization of the multi-pulse scenario with a pulse width of 0.12, and all other parameters at baseline levels (see Table 2). Left panel: data including sample means. Right panel: corresponding p-value functions with threshold level 0.05.

4.2 Performance measures

Performance of the IWT, SPM, SnPM and BH methods were evaluated using two different measures: sensitivity and FWER. Sensitivity was estimated as the average proportion of the domain where the signal was correctly identified (i.e., proportion of $\mathcal{D}_1$ where H_0 was rejected). FWER was estimated as the

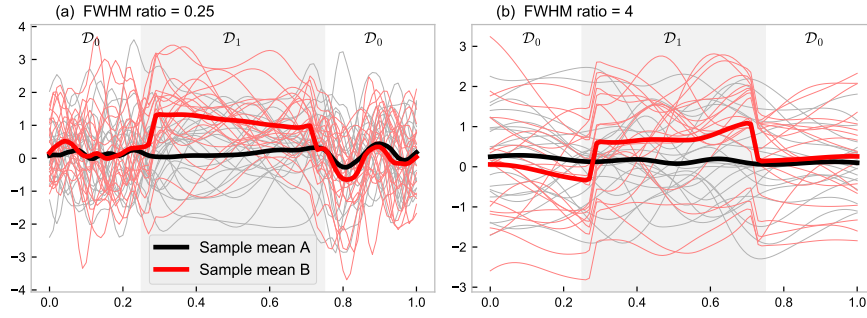


Fig. 8 Example realizations of non-uniformly smooth datasets with FWHM ratios = 0.25 and 4, where the ratio value represents the smoothness of $\mathcal{D}_0$ relative to $\mathcal{D}_1$, and where the smoothness in $\mathcal{D}_1$ is FWHM=0.2, and all other parameters are at baseline levels.

proportion of simulations where the null hypothesis was wrongly rejected for at least one point of the domain (i.e., rejected for at least one point in $\mathcal{D}_0$). As a reference, FWER was additionally quantified for the case of null signal (i.e., signal amplitude = 0), where $\mathcal{D}_0$ spanned the entire domain; note that sensitivity is undefined for this null signal case. Last, the average computational duration was calculated for each procedure over 10,000 simulation iterations of the null signal case.

4.3 Results

Excluding unadjusted results, the estimated sensitivity was generally highest for BH and the lowest for IWT, with SPM and SnPM exhibiting intermediate sensitivity (Fig.9). Exceptions to this trend occurred near the extremes of some parameter ranges including, for example: large sample size imbalance, for which SnPM sensitivity became poorer than IWT sensitivity (Fig.9b) and very rough domains (FWM < 0.1), for which IWT sensitivity was greater than SPM and SnPM's (Fig.9f).

Sensitivity increased with sample size, sample size balance, signal amplitude and signal width, and the relation amongst all methods' sensitivities was largely unaffected by these parameters (Fig.9a-e). An exception to this pattern was observed for very narrow (<0.1) and very broad signal widths (>0.5), for which BH and IWT results respectively converged to the SPM/SnPM results (Fig.9d,e). A similar convergence between BH and SPM/SnPM was observed for very smooth domains (FWHM>0.5) (Fig.9f). A similar convergence between BH and IWT results was observed when the $\mathcal{D}_0:\mathcal{D}_1$ variance ratio was small (< 0.5) (Fig.9h).

Skewness was found to increase sensitivity in all procedures (Fig.9g). Additionally, skewness did not markedly affect the relative sensitivities of the three procedures.

FWER estimates were relatively stable, with IWT, SPM and SnPM exhibiting FWER values close to $\alpha = 0.05$ for all simulations (Fig.10). The BH method's FWER results were contrastingly parameter-dependent and generally much greater than α . Most notably, BH's estimated FWER was very large (> 0.3) for highly rough domains (Fig.10f,g).

FDR estimates were relatively stable for the BH method, with FDR values close to $\alpha = 0.05$ for all simulations (Fig.11). The IWT, SPM and SnPM methods tended to exhibit smaller FDR values, except in extreme cases; in the most extreme case, the SnPM method exhibited an FDR of approximately 0.10 in the highly unequal variance scenario (Fig.11i). The unadjusted approach exhibited considerably greater FDR, with FDR values of approximately 0.10 in most cases, but exceeding 0.30 in extreme cases including very narrow signals (Fig.11d) and highly unequal variance (Fig.11i).

Computational durations (means $\pm$ standard deviations over 10,000 simulation iterations of the baseline scenario) were: IWT = 4.040 ± 0.190 s, SPM = 0.064 ± 0.006 s, SnPM = 0.162 ± 0.016 s, and BH = 0.0031 ± 0.0005 s.

Table 3 Family wise error rate (FWER) results for the case of null signal (i.e., $\mathcal{D}_1$ is the null interval). IWT, SPM and SnPM control FWER for this case.

Procedure	FWER
Unadjusted	0.394
IWT	0.005
SPM	0.051
SnPM	0.052
BH	0.027

5 Discussion

The primary aims of this paper were to compare four different functional data analysis procedures in the context of two-sample comparisons of sports datasets, and to conduct a relatively thorough ten-parameter simulation study which mimicked some of the key features of the experimental datasets. Experimental data analyses revealed a variety of similarities across datasets and procedures, including especially generally lower adjusted p-values with BH and IWT than with SPM and SnPM, as well as larger sub- α domain regions. Null hypothesis rejection decisions (i.e., whether the minimum p-value was less than α) were largely identical across datasets and procedures.

Simulation results showed that BH had the greatest sensitivity, but also that it exhibited relatively large FWER (generally $\text{FWER} > 0.1$), especially for the extremes of the investigated smoothness parameters ($\text{FWER} > 0.3$) (Fig.10f,g). SPM and SnPM exhibited greater sensitivity than IWT in most investigated scenarios, and IWT exhibited greater sensitivity only when (i) the data were highly rough ($\text{FWHM} < 0.1$) (Fig.10f), (ii) $\mathcal{D}_0$ data were more than twice as rough as $\mathcal{D}_1$ data (Fig.10g), or (iii) variance in $\mathcal{D}_0$ was smaller

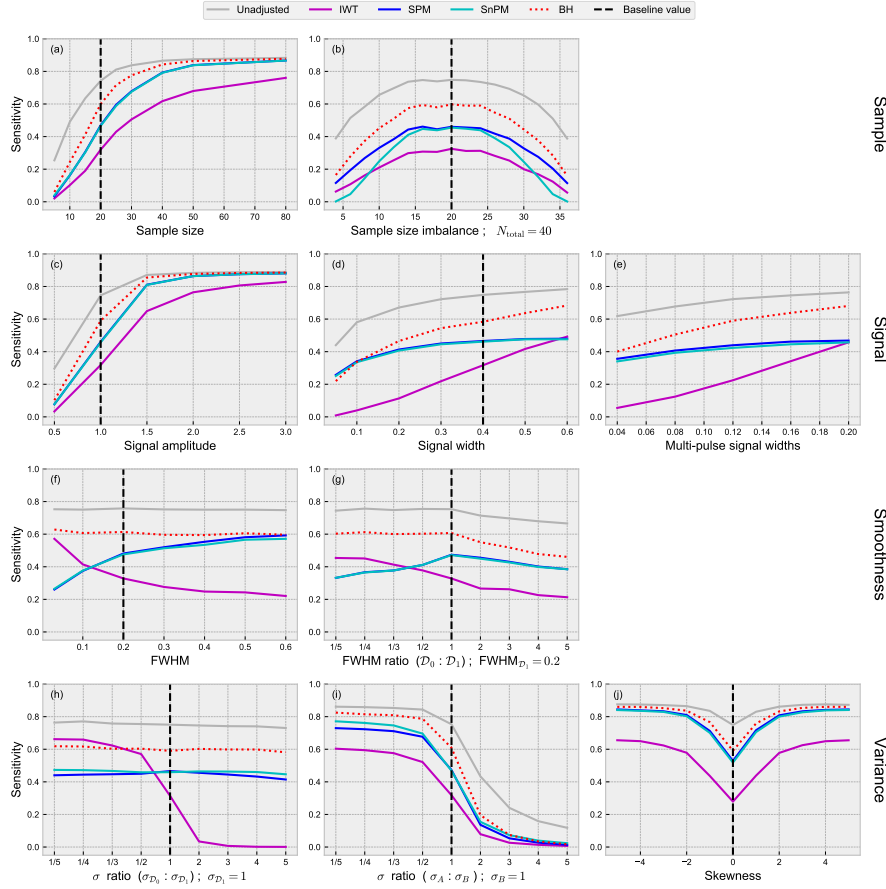


Fig. 9 Estimated sensitivity results for each parameter. Rows represent the parameter categories listed in Table 2. All panels except (e) involve the baseline scenario; note that baseline results are equivalent across panels, within numerical tolerance. See Table 2, Fig. 5 and §4.1 for descriptions of terms.

than in $\mathcal{D}_1$ (Fig. 10h). Since apparently greater IWT sensitivity was observed in the experimental datasets, these simulation results suggest either that the experimental data followed one of these three characteristics, or that these simulations probed an insufficient portion of the conceivable parameter space.

From qualitative considerations of the experimental results (Figs. 1-4), their underlying sample variances (Fig. 12) and smoothness nonuniformity (Fig. 13), it appears that large variance in large-effect regions is a plausible explanation for the higher IWT sensitivity in the Kautz et al. (1991) dataset (Fig. 1). Similarly, high smoothness in large-effect regions appears to be a plausible explanation for increased IWT sensitivity in the Neptune et al. (1999) and Dorn et al. (2012) datasets (Figs. 2, 4). Of the two remaining two datasets from Besier et al. (2009) (Fig. 3), one (top panel) is consistent with a generally

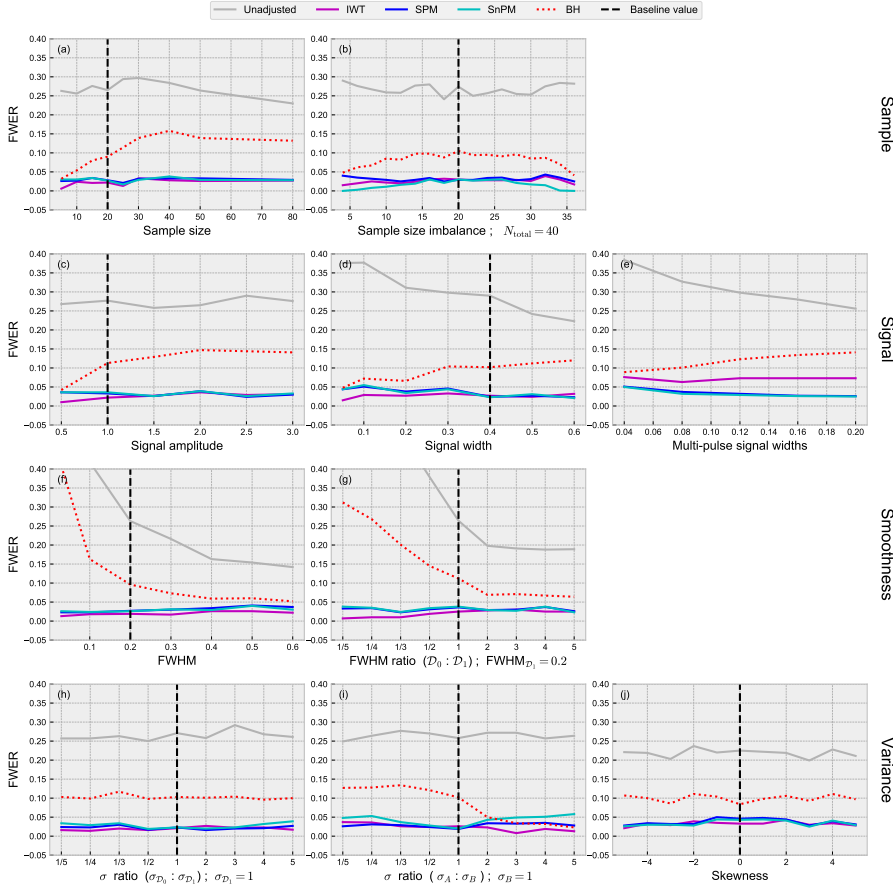


Fig. 10 Estimated FWER results for each parameter. Data presented as in Fig.9.

higher sensitivity in SPM and SnPM vs. IWT, and the other (bottom panel) involves an apparently very broad large-effect region. These results generally agree with the wide-signal simulation results in Fig.9e, and are thus indicative of convergence between IWT and SPM/SnPM for very broad signals.

Since the simulation scenarios can qualitatively explain the main features of the experimental data results, we conclude that (i) the investigated simulation parameters appear to be sufficient to explain a variety of effects in representative biomechanics datasets, and that (ii) nonuniform smoothness and nonstationarity appear to be inherent features of some biomechanics datasets, which implies that IWT might be more sensitive than SPM/SnPM for these datasets if the nonuniform smoothness and/or nonuniformity profiles coincide with large-effect regions. Regardless, since true effects are unknown for experimental datasets, it is generally not possible to conclude whether one procedure would be more sensitive than other. Instead we can generally only recommend the following: (i) apply all of the procedures to a given dataset,

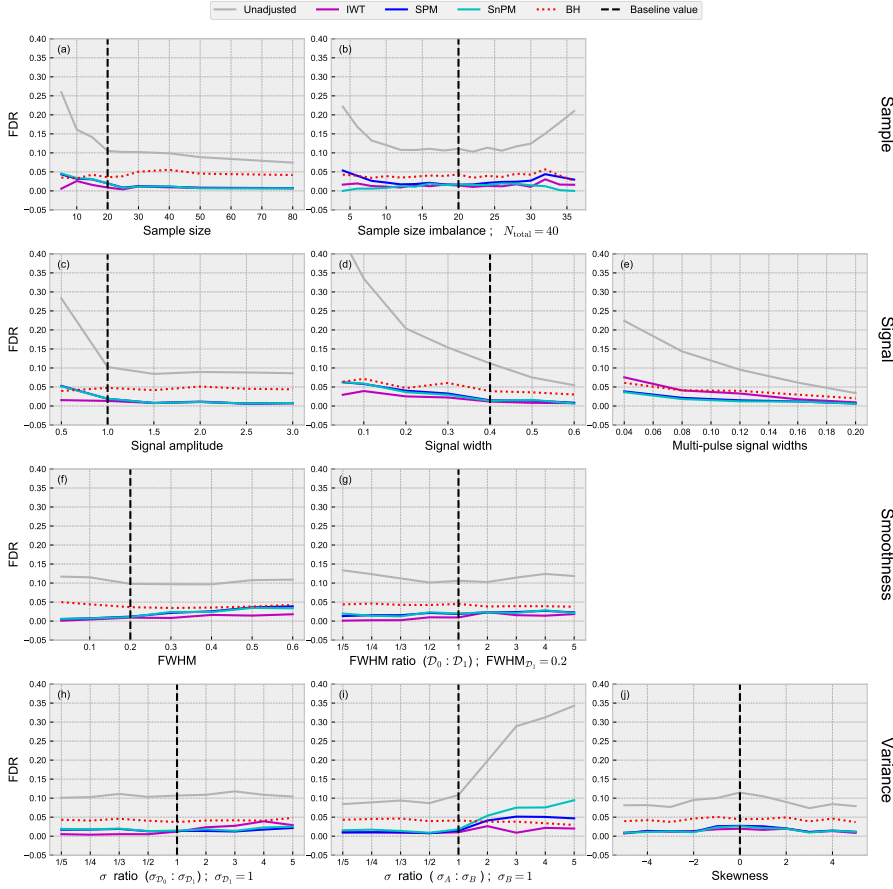


Fig. 11 Estimated FDR results for each parameter. Data presented as in Fig.9.

then (ii) aim to explain differences amongst the various procedures' results in terms of the currently demonstrated effects of nonuniform smoothness and/or nonstationarity.

The main reasons we selected the four current techniques (IWT, SPM, SnPM and IWT) are that (i) they can all be used as domain selection techniques, and (ii) they have all been used in the recent sports science literature. Other methods for domain selection have been proposed in the FDA-literature (e.g. [Vsevolozhskaya et al., 2015](#)), but these methods do not seem to be regularly used in the analysis of sports related data.

The current BH results agree with a recent comparison to SPM which showed that, while the methods yield similar results for a wide variety of sports-relevant experimental datasets, the methods' results can differ considerably for specific datasets ([Naouma and Pataky, 2019](#)). These basic findings have since been clarified for a wider variety of simulated and experimental datasets ([Olsen et al., 2021](#)). The main indication is that BH sensitivity in-

creases as signal amplitude increases, especially as signal width increases. This well-known sensitivity property of BH is likely superior to all domain selection methods including IWT and SPM/SnPM, and it is consistent with the fact that FDR control is weaker than FWER control.

This study’s simulations were limited by a variety of factors. Simulation parameters were varied only one at a time, relative to a baseline scenario (Fig.6). No systematic exploration of parameter space was conducted because a systematic search of the ten-dimensional parameter space was computationally infeasible. We prioritized a parameter-by-parameter exploration over an exhaustive parameter space search because this was the first systematic comparison amongst the four techniques, and we wished to provide an overview of how individual parameters can affect the various approaches’ results. The baseline scenario used here was chosen to generally represent the characteristics of the experimental datasets. Still, to fully understand the methods’ similarities and differences it would be of interest to vary multiple parameters simultaneously.

Another limitation was that the baseline scenario was restricted to the case of a single domain $\mathcal{D}_1$, with the exception of the ‘multi-pulse’ scenario (Fig.7), for which $\mathcal{D}_1$ was a union of three intervals. Sensitivity and FWER were estimated for different sizes of these multi-pulse intervals. A natural extension would be to consider other situations where $\mathcal{D}_1$ is decomposed into multiple disjoint parts/intervals.

A third limitation was that the non-Gaussian distributions considered here were only mildly non-Gaussian, and were qualitatively rather symmetrical (see Appendix, Fig.15). While the maximum absolute skewness parameter value was sufficient to yield significance in distribution tests (results not shown), it may inadequately reflect the degree of non-normality that is present in specific sports biomechanics datasets (Fig.14). It would be of interest to systematically explore a wider variety of non-Gaussian error distributions.

A fourth limitation is that our implementation of distribution skewed data did not precisely control smoothness of the domain, and this caused the baseline results (Fig.9j) to be slightly different than the baseline scenario in the other panels. Since the same general trends are expected with skewed data, regardless of smoothness control precisions, we regard this as a minor point, and irrelevant to this paper’s conclusions.

A minor, but practical limitation of IWT is its computational speed. It was approximately 60,25, and 1300 times slower than SPM, SnPM and BH, respectively. It is possible that these speed differences are partially attributable to computational properties of R vs. Python. This is not a serious limitation for offline analysis, but may make it inconvenient or impractical in time-sensitive applications like athlete feedback. Additionally, IWT is not easily generalized to higher dimensions, since the analysis is performed on intervals of the domain. For higher dimensions, it is less trivial how to construct “intervals”, and the computational time would be extensive.

We note that SPM’s adjusted p-values are asymptotically accurate as the test statistic value increases, and that large p-values (>0.5) are generally in-

valid (Friston et al., 2007). One reason is that topological (i.e., EEC) behavior stabilizes as the test statistic increases. That is, thresholded functions exhibit stabler topological behavior when thresholded at high values vs. low values. As this study was primarily interested in low p-values, this was not considered an important limitation to this study, nor relevant to its conclusions. We therefore presented all SPM-adjusted p-values, regardless of known inaccuracies. We also note that, since SnPM’s distributions are data-derived using permutation, this accuracy limitation does not apply to SnPM.

One final, and minor limitation, is that the Neptune et al. (1999) dataset had an angular dependent variable (Fig.2). For this dataset, circular procedures would have been more appropriate than the Euclidean procedures adopted in this paper. As far as we are aware, no IWT, SPM, SnPM or BH software package supports circular data analysis. Additionally, the angular displacements were not very large (approximately 40 deg, Fig.2), implying that circular results would not be expected to be very different from the current linear approximations. Regardless, the main point of this paper was to consider the differences amongst the different domain selection procedures, so the angular nature of this DV was considered to be irrelevant to this paper’s conclusions.

In summary, this study has shown that IWT, SPM, SnPM and BH performed similarly for a variety of experimental datasets and in a variety of experimental scenarios. BH tended to be the most sensitive technique, at the expense of a relatively high FWER of 0.1 or greater. IWT tended to be more sensitive than SPM and SnPM for experimental dataset analysis, and this difference in sensitivity may be explained by certain dataset characteristics including nonstationarity and nonuniform smoothness. In terms of the primary hypothesis testing result — the H_0 rejection decision — the relatively small differences amongst the four techniques imply that there is likely little practical difference amongst these techniques when applied over many datasets. Nevertheless, since dataset characteristics can influence, and even reverse their relevant sensitivities, we recommend that users (i) apply all procedures wherever possible, (ii) report all dataset characteristics, especially nonstationarity, smoothness, and smoothness nonuniformity, and (iii) insofar as the techniques disagree, interpret all results with caution. Recommendation ‘(ii)’ above (reporting dataset characteristics) will be useful for future work, to bring consistency and transparency to the literature involving domain selection techniques and their results.

Acknowledgements The authors acknowledge support from The Swedish Scientific Research Council (Grant Nos. 2016-02763 and 340-2013-5203).

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Abramowicz, K., Häger, C. K., Pini, A., Schelin, L., Sjöstedt de Luna, S. and Vantini, S. (2018) Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. *Scandinavian Journal of Statistics*, **45**, 1036–1061.
- Adler, R. and Hasofer, A. (1976) Level Crossings for Random Fields. *The Annals of Probability*, **4**, 1–12.
- Adler, R. J. and Taylor, J. E. (2007) *Random Fields and Geometry*. Springer-Verlag.
- Anaconda (2020) Anaconda Software Distribution version 3-6.10. URL: <https://anaconda.com>.
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, **57**, 289–300.
- Besier, T. F., Fredericson, M., Gold, G. E., Beaupré, G. S. and Delp, S. L. (2009) Knee muscle forces during walking and running in patellofemoral pain patients and pain-free controls. *Journal of biomechanics*, **42**, 898–905.
- Chumbley, J. R. and Friston, K. J. (2009) False discovery rate revisited: Fdr and topological inference using gaussian random fields. *Neuroimage*, **44**, 62–70.
- Dannenmaier, J., Kaltenbach, C., Kölle, T. and Krischak, G. (2020) Application of functional data analysis to explore movements: walking, running and jumping—a systematic review. *Gait & Posture*, **77**, 182–189.
- Dorn, T. W., Schache, A. G. and Pandy, M. G. (2012) Muscular strategy shift in human running: dependence of running speed on hip and ankle muscle performance. *Journal of Experimental Biology*, **215**, 1944–1956.
- Friston, K., Holmes, A., Worsley, K., Poline, J., Frith, C. and Frackowiak, R. (1995) Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, **2**, 189–210.
- Friston, K. J., Ashburner, J. T., Kiebel, S. J., Nichols, T. E. and Penny, W. D. (2007) *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. London: Elsevier.
- Genovese, C. R., Lazar, N. A. and Nichols, T. (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, **15**, 870–878.
- Hébert-Losier, K., Schelin, L., Tengman, E., Strong, A. and Häger, C. K. (2018) Curve analyses reveal altered knee, hip, and trunk kinematics during drop-jumps long after anterior cruciate ligament rupture. *The Knee*, **25**, 226–239.
- Kautz, S. A., Feltner, M. E., Coyle, E. F. and Baylor, A. M. (1991) The pedaling technique of elite endurance cyclists: changes with increasing workload at constant cadence. *Journal of Applied Biomechanics*, **7**, 29–53.
- Kiebel, S., Poline, J., Friston, K., Holmes, A. and Worsley, K. (1999) Robust Smoothness Estimation in Statistical Parametric Maps Using Standardized Residuals from the General Linear Model. *NeuroImage*, **10**, 756–766.

- Markström, J. L., Grip, H., Schelin, L. and Häger, C. K. (2019) Dynamic knee control and movement strategies in athletes and non-athletes in side hops: Implications for knee injury. *Scandinavian journal of medicine & science in sports*, **29**, 1181–1189.
- Naouma, H. and Pataky, T. C. (2019) A comparison of random-field-theory and false-discovery-rate inference results in the analysis of registered one-dimensional biomechanical datasets. *PeerJ*, **7**, e8189.
- Neptune, R. R., Wright, I. C. and Van Den Bogert, A. J. (1999) Muscle coordination and function during cutting movements. *Medicine and science in sports and exercise*, **31**, 294–302.
- Nichols, T. and Holmes, A. (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, **15**, 1–25.
- O’Hagan, A. and Leonard, T. (1976) Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, **63**, 201–203.
- Olsen, N. L., Pini, A. and Vantini, S. (2021) False discovery rate for functional data. *Test*, in press, doi: 10.1007/s11749-020-00751-x.
- Pataky, T. C. (2012a) One-dimensional statistical parametric mapping in python. *Computer Methods in Biomechanics and Biomedical Engineering*, **15**, 295–301.
- (2012b) Spatial resolution in plantar pressure measurement revisited. *Journal of Biomechanics*, **45**, 2116–2124.
- (2016) rft1d: Smooth One-Dimensional Random Field Upcrossing Probabilities in Python. *Journal of Statistical Software*, **71**, 1–22.
- Pataky, T. C., Robinson, M. A. and Vanrenterghem, J. (2013) Vector field statistical analysis of kinematic and force trajectories. *Journal of biomechanics*, **46**, 2394–2401.
- Pataky, T. C., Vanrenterghem, J. and Robinson, M. A. (2015) Zero- vs. one-dimensional, parametric vs. non-parametric, and confidence interval vs. hypothesis testing procedures in one-dimensional biomechanical trajectory analysis. *Journal of Biomechanics*, **48**, 1277–1285.
- Pataky, T. C., Vanrenterghem, J., Robinson, M. A. and Liebl, D. (2019) On the validity of statistical parametric mapping for nonuniformly and heterogeneously smooth one-dimensional biomechanical data. *Journal of biomechanics*, **91**, 114–123.
- Pini, A., Spreafico, L., Vantini, S. and Vietti, A. (2019) Multi-aspect local inference for functional data: Analysis of ultrasound tongue profiles. *Journal of Multivariate Analysis*, **170**, 162–185.
- Pini, A. and Vantini, S. (2017) Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, **29**, 407–424.
- R Core Team (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.
- Ramsay, J. O. and Silverman, B. W. (2005) *Functional data analysis*. Springer, New York.

- Richter, C., Marshall, B., Moran, K. et al. (2014) Comparison of discrete-point vs. dimensionality-reduction techniques for describing performance-related aspects of maximal vertical jumping. *Journal of biomechanics*, **47**, 3012–3017.
- van Rossum, G. (2014) The python library reference release 3.7.1. URL: <https://docs.python.org/3/library/>.
- Shannon, C. E. (1949) Communication in the presence of noise. *Proceedings of the IRE*, **37**, 10–21.
- Vsevolozhskaya, O. A., Greenwood, M. C., Powell, S. L. and Zaykin, D. V. (2015) Resampling-based multiple comparison procedure with application to point-wise testing with functional data. *Environmental and ecological statistics*, **22**, 45–59.
- Warmenhoven, J., Harrison, A., Robinson, M. A., Vanrenterghem, J., Bargary, N., Smith, R., Cobley, S., Draper, C., Donnelly, C. and Pataky, T. (2018) A force profile analysis comparison between functional data analysis, statistical parametric mapping and statistical non-parametric mapping in on-water single sculling. *Journal of Science and Medicine in Sport*, **21**, 1100–1105.
- Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J. and Evans, A. C. (1996) A Unified Statistical Approach for Determining Significant Signals in Images of Cerebral Activation. *Human Brain Mapping*, **4**, 58–73.
- Yang, J. and Ren, P. (2019) Bfda: A matlab toolbox for bayesian functional data analysis. *Journal of Statistical Software, Articles*, **89**, 1–21. URL: <https://www.jstatsoft.org/v089/i02>.

Appendix

Additional figures for the manuscript.

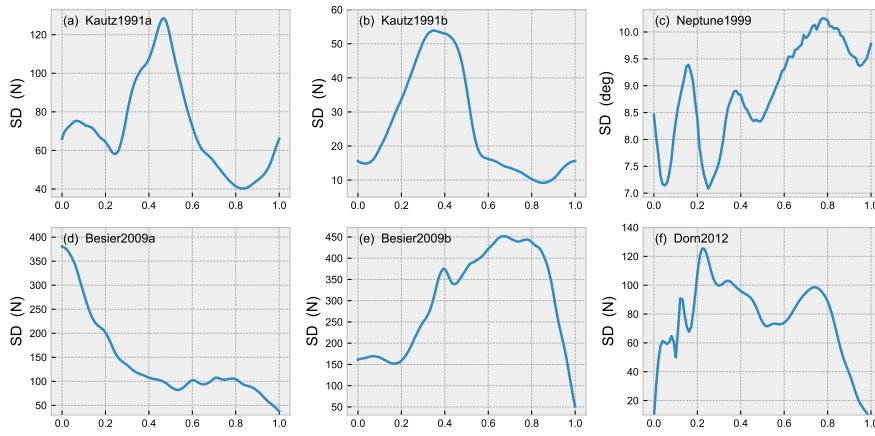


Fig. 12 Estimated sample and pooled standard deviations (SD) in the experimental datasets.

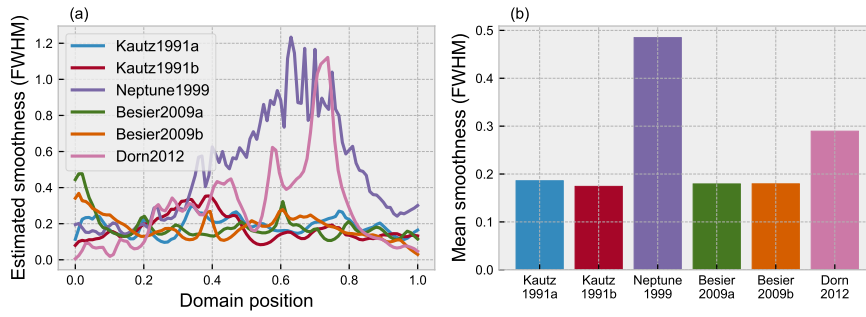


Fig. 13 Estimated smoothness (FWHM) in the experimental datasets (units: relative to domain size). (a) FWHM functions. (b) Mean FWHM values across the domain.

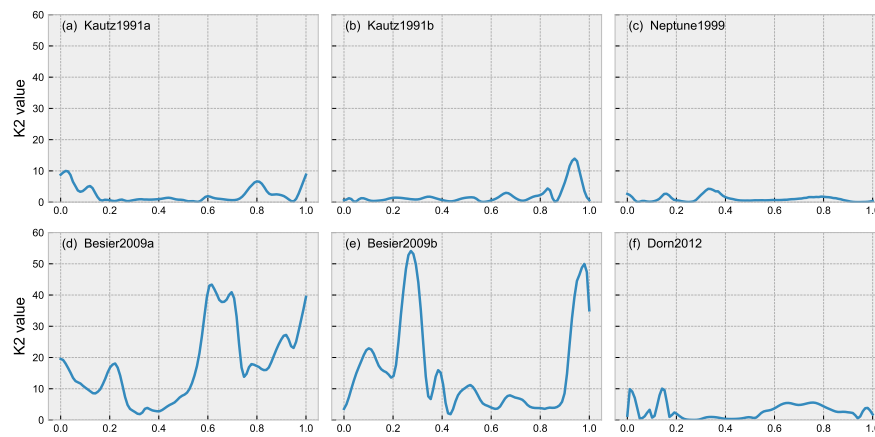


Fig. 14 D'Agostino-Pearson K2 (non-normality) values. Large K2 values indicate strong deviation from normality.

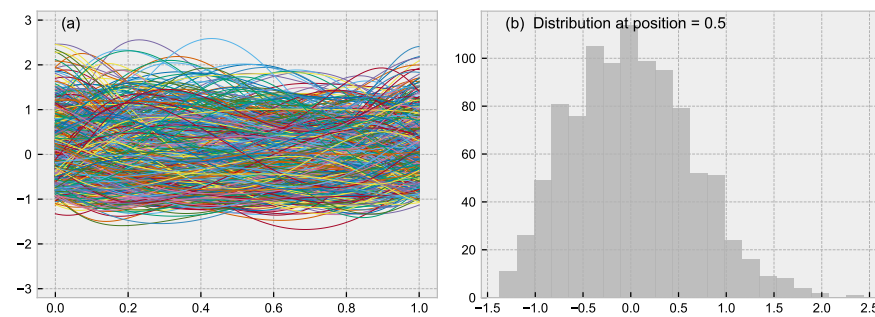


Fig. 15 Example skew-normal noise, with skew parameter $\alpha = +5$ (skew normal distribution) (O'Hagan and Leonard, 1976). A sample size of 1000 is presented for distributional clarity.